Stats 4CI3 - Assignment 1

Tommy Flynn (400121496)

January 27, 2021

Question 1:

a) Here is the requested vector.

```
v = -15:15
v
```

```
[1] -15 -14 -13 -12 -11 -10  -9  -8  -7  -6  -5
[12]  -4  -3  -2  -1   0   1   2   3   4   5   6
[23]   7   8   9  10  11  12  13  14  15
```

b) The requested sequences.

```
v1 = seq(1,5,1)
v2 = seq(3,30,3)
v1
v2
```

```
[1] 1 2 3 4 5
[1]  3  6  9 12 15 18 21 24 27 30
```

c) Here are the indices of $v2$ greater than 10.

```
which(v2>10)
```

```
[1] 4  5  6  7  8  9 10
```

d) The requested plot.

```
curve(sin(x),0,8*pi)
```

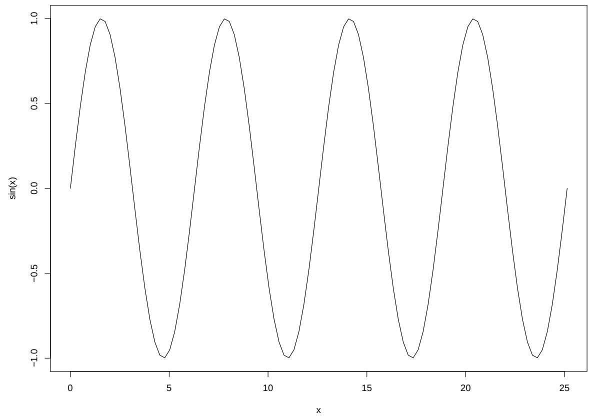

Figure 1: Sin(x) on $(0, 8\pi)$

e) Lastly, a while loop to sum squares up to 10.

```
i = 1
s = 0
while(i<11){
  s = s + i^2
  i = i + 1
}
s
```

```
[1] 385
```

Question 2:

a) I used the dataset 'Forest Fires' from https://archive.ics.uci.edu/ml/datasets/Forest+Fires

b) Here is an explanation of the variables which was obtained from the same URL.

```
1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84
(this output variable is very skewed towards 0.0, thus it may make
sense to model with the logarithm transform).
```

And, here is a numerical summary using R.

```r
ff = read.csv("/Users/tommyflynnrogers.com/Desktop/forestfires.csv")
summary(ff)
```

```
      X               Y             month        day         FFMC
 Min.   :1.000   Min.   :2.0   aug    :184   fri:85   Min.   :18.70
 1st Qu.:3.000   1st Qu.:4.0   sep    :172   mon:74   1st Qu.:90.20
 Median :4.000   Median :4.0   mar    : 54   sat:84   Median :91.60
 Mean   :4.669   Mean   :4.3   jul    : 32   sun:95   Mean   :90.64
 3rd Qu.:7.000   3rd Qu.:5.0   feb    : 20   thu:61   3rd Qu.:92.90
 Max.   :9.000   Max.   :9.0   jun    : 17   tue:64   Max.   :96.20
                               (Other): 38   wed:54
      DMC              DC              ISI              temp             RH
 Min.   :  1.1   Min.   :  7.9   Min.   : 0.000   Min.   : 2.20   Min.   : 15.00
 1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500   1st Qu.:15.50   1st Qu.: 33.00
 Median :108.3   Median :664.2   Median : 8.400   Median :19.30   Median : 42.00
 Mean   :110.9   Mean   :547.9   Mean   : 9.022   Mean   :18.89   Mean   : 44.29
 3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800   3rd Qu.:22.80   3rd Qu.: 53.00
 Max.   :291.3   Max.   :860.6   Max.   :56.100   Max.   :33.30   Max.   :100.00

      wind             rain              area
 Min.   :0.400   Min.   :0.00000   Min.   :   0.00
 1st Qu.:2.700   1st Qu.:0.00000   1st Qu.:   0.00
```

```
Median :4.000   Median :0.00000   Median :   0.52
Mean   :4.018   Mean   :0.02166   Mean   :  12.85
3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.:   6.57
Max.   :9.400   Max.   :6.40000   Max.   :1090.84
```

b) Here is a histogram and scatter plot.

```
hist(ff$DMC, xlab="Duff Moisture Code (DMC)", ylab="Count", main="Moisture Histogram")
plot(ff$temp,ff$DMC, xlab="Temperature in Celsius", ylab="Duff Moisture Code (DMC)",
     main="Temperature Moisture Scatter")
```

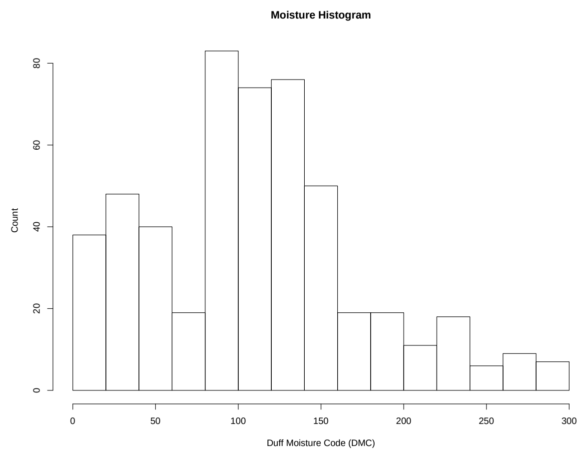

Figure 2: Histogram

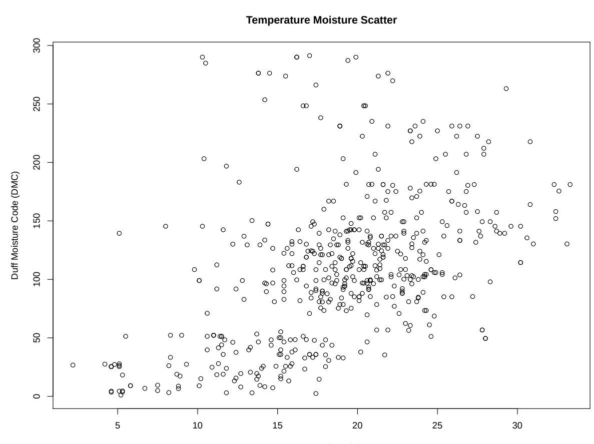

Figure 3: Scatter

d) Now on the same plot.

```
par(mfrow=c(1,2))
hist(ff$DMC, xlab="Duff␣Moisture␣Code␣(DMC)", ylab="Count", main="Moisture␣Histogram")
plot(ff$temp,ff$DMC, xlab="Temperature␣in␣Celsius", ylab="Duff␣Moisture␣Code␣(DMC)",
     main="Temperature␣Moisture␣Scatter")
```

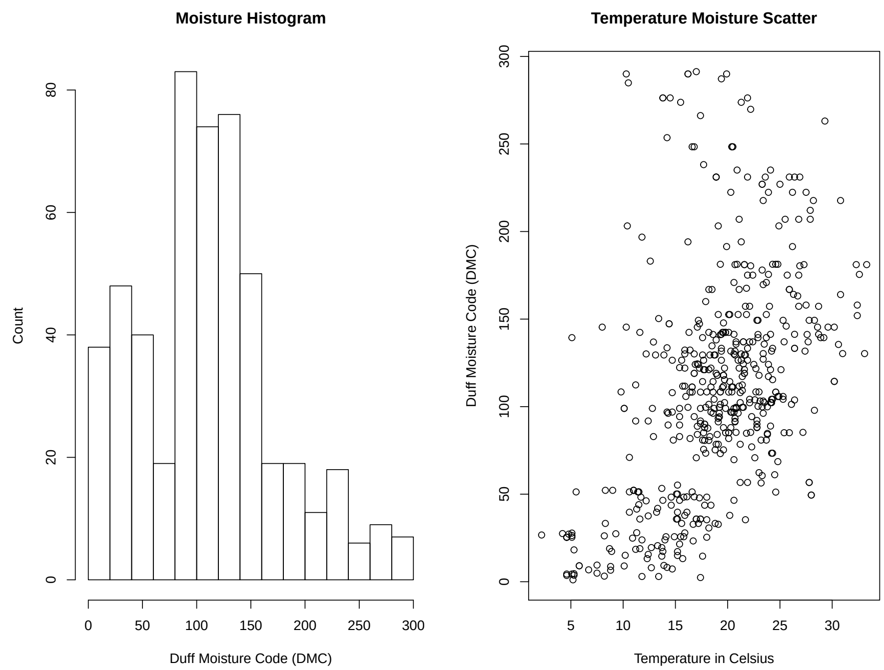

Figure 4: Histogram and Scatter

Question 3:

a) We are to test the null hypothesis that the mean of population x is equal to the mean of population y, that is, $H_0 : \mu_x = \mu_y$ v.s. $H_a : \mu_x \neq \mu_y$. Here is the test statistic for the proposed hypothesis test

$$T = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{1/m + 1/n}} \sim t_{m+n-2}$$

where s_p is the pooled standard deviation. Now the code.

```
set.seed(6573)
x = rnorm(50,80,20)
y = rnorm(50,80,20)

m = length(x)
n = length(y)
sp=sqrt(((m-1)*sd(x)^2+(n-1)*sd(y)^2)/(m+n-2))
t.stat=(mean(x)-mean(y))/(sp*sqrt(1/m+1/n))
t.stat
```

```
[1] -0.2175905
```

b) Here is the critical value t_{m+n-2}.

```
alpha=0.05
cv=qt(1-alpha/2,n+m-2)
cv
```

```
[1] 1.984467
```

c) We fail to reject the null hypothesis since the test statistic is absolutely less than the critical value i.e. $|-0.2175905| < 1.984467$.

Question 4:

a) We are to perform a lower tail test of the population mean, that is, $H_0 : \mu_0 \leq \mu$ v.s. $H_a : \mu_0 > \mu$. Here is the test statistic for for the proposed hypothesis test.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Now the code.

```
z = (7900 - 8000)/(100/sqrt(50))
z
```

```
[1] -7.071068
```

b) Here is the critical value.

```
cv2 = qnorm(1 - alpha)
cv2
```

```
[1] 1.644854
```

c) We reject the null hypothesis since our test statistic is less than the critical value i.e. $-7.071068 < 1.644854$.